Enhancing Intelligent Customer Service Systems with Modular LLM Architecture

Mingyu Zhang*
Xi'an Jiaotong-Liverpool University, Suzhou, China
**Corresponding Author*

Abstract: In the contemporary business landscape, rapid advances in artificial intelligence (AI) and machine learning have led to the emergence of intelligent customer service (ICS) as a transformative solution. ICS leverages advanced technologies such as natural language processing (NLP), machine learning algorithms, and automation to provide personalized, efficient customer support. This paper presents an in-depth exploration of ICS, examining its key components, benefits, challenges, and future trends. It reviews existing literature on AI-driven customer service, analyzes the impact of intelligent systems on customer satisfaction and operational efficiency, and discusses the ethical and privacy considerations associated with deploying these technologies. By synthesizing the latest research and industry practices, the study provides insights for businesses seeking to enhance their customer service capabilities through intelligent solutions.

Keywords: LLM; RAG; Human-Computer Collaboration; Customer Service

1. Introduction

The rapid growth of business and commerce has significantly increased the demand for customer service. Traditional systems often struggle to handle large-scale interactions, leading to high operational costs and extensive training requirements [1]. Recent advancements in artificial intelligence (AI) and large language models (LLMs) offer a more efficient alternative [2].

LLMs exhibit strong conversational capabilities, enabling effective multilingual communication, sentiment analysis, and accurate query resolution (Brown et al., 2020). Enhanced by developments in natural language processing (NLP), LLMs can process complex customer inquiries with high precision [3], and concurrently serve large volumes of users—a scalability unattainable by human agents. Furthermore, LLMs demonstrate adaptability to evolving business contexts and market demands.

Human–computer interaction (HCI) plays a critical role in optimizing LLMs. Effective prompt engineering provides clear instructions and context, improving the naturalness and accuracy of generated responses. Techniques such as role-playing and emotional prompts enhance content richness, while specifying sources reduces misinformation. Evaluation metrics like BLEU, ROUGE, and BERTScore, along with reference-free metrics, support quality assessment. Self-evaluation capabilities and human-in-the-loop (HITL) review processes, particularly in sensitive domains such as law and healthcare, further ensure content reliability. HCI also enhances customer experience by enabling LLMs to interpret emotions and intent, balancing task-oriented and open-domain interactions through user-centered design and feedback [4].

Prompt engineering is integral to this process. Structured prompts significantly improve output relevance and precision [5]. Retrieval-Augmented Generation (RAG) integrates external knowledge into prompts, reducing hallucination and enhancing factual accuracy. By clarifying instructions, incorporating context, and dynamically refining content, RAG supports more accurate and professional customer interactions [6].

In summary, LLMs, augmented by prompt engineering and RAG, offer substantial improvements over traditional customer service. They provide scalable, cost-effective, and multilingual support with high accuracy and personalization, effectively meeting diverse customer needs.

2. Related Work

The application of AI in customer service has garnered substantial research attention. Early

efforts centered on rule-based chatbots and virtual assistants, which were constrained in their ability to address complex queries. The evolution of NLP and machine learning has led to more sophisticated systems capable of managing diverse customer service tasks [7].

Gentch (2018) conducted a systematic review of AI applications in customer service, noting the transition from rule-based frameworks to NLP-driven solutions. With the rise of LLMs, intelligent question-answering systems have become central to various domains, including intelligent customer support, knowledge retrieval, and online education. LLMs are also effective in solving mathematical problems, further showcasing their versatility[8].

The integration of LLMs into customer service operations has proven transformative, significantly improving response efficiency and customer satisfaction. This section reviews current research and practical implementations of LLMs in customer service, emphasizing their role in advancing intelligent service systems [9].

3. Methodology

This study introduces an advanced modular architecture designed for intelligent customer service systems, powered by Large Language Models (LLMs). The proposed framework is structured into four integral components, which is shown in Fig. 1, each addressing distinct challenges in the realm of customer service automation. These components work synergistically to enhance the system's performance in providing accurate, efficient, and contextually relevant assistance. Below is a detailed description of each module:

3.1 User Input Processing Module

The first step in the interaction process involves user queries, which may often be ambiguous, incomplete, or contain extraneous information. To address these issues, this module employs sophisticated natural language processing (NLP) techniques to pre-process and refine the raw inputs. The goal of this stage is to eliminate noise, identify key intent, and resolve any inherent ambiguities, thereby enabling the system to better understand user needs. Through this enhancement, the system can generate more accurate, relevant, and context-sensitive responses, ensuring that the interaction is both efficient and meaningful.

3.2. Information Retrieval Module

At the core of the system's intelligence is the Information Retrieval (IR) module. This component utilizes a state-of-the-art dense retrieval mechanism to search and extract pertinent data from an expansive knowledge base. By identifying semantically similar information, the system ensures that the responses are not only factually correct but also tailored to the specific context of the user's inquiry. This approach optimizes the retrieval of both specific product information and generalized knowledge, making it an indispensable tool for delivering accurate, data-driven answers. The integration of semantic search ensures that even complex queries are understood and addressed with precision.

3.3. Response Generation Module

Once the relevant information is retrieved, the Generation module synthesizes the user input with the newly acquired knowledge to craft coherent and contextually aware responses. The system dynamically integrates the user’s query with auxiliary content, enabling it to generate answers that are not only accurate but also comprehensive. This module leverages advanced language models to ensure that responses are fluent, contextually appropriate, and rich in detail. Additionally, it is designed to adapt to the tone and complexity of the input, ensuring that all generated content aligns with the user's expectations and communication style.

3.4 Human-Computer Interaction Interface Module

The final component, the Human-Computer Interaction (HCI) module, is responsible for facilitating smooth and intuitive interactions between the user and the intelligent system. The user interface (UI) is designed to accommodate natural language inputs, as well as multimodal communication, including image uploads and possibly other media types. This module ensures that the system is highly interactive, accessible, and easy to use, fostering a user-friendly experience. By supporting a variety of interaction modalities, it enhances the versatility of the system, allowing it to serve a broader range of user needs and contexts. Through this seamless interface, users can effortlessly communicate with the system, ensuring that the overall experience is both effective and engaging.

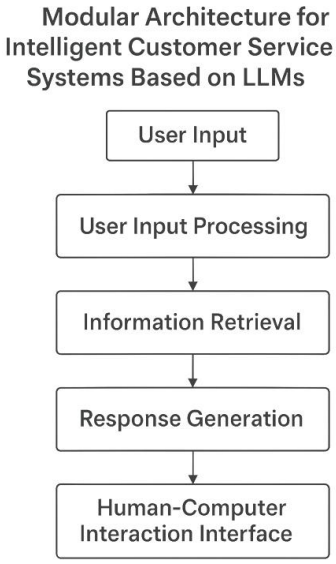

Figure 1. The Framework of Our Work

4. Experiment

The experiment is designed to assess the performance of the system across each module and the overall integration. We will evaluate metrics such as accuracy, response time, user satisfaction, and the relevance of responses. The experiment will involve multiple iterations to ensure the robustness of the system's performance.

4.1 Evaluation Matric

The evaluation matrix used in this study scores the system's performance across four key metrics: Accuracy, Response Time, User Satisfaction, and Relevance. Accuracy is weighted at 40%, reflecting the importance of delivering correct responses based on expected outcomes. Response Time, which measures the speed at which the system generates and delivers a response, is given a weight of 25%, highlighting the need for timely interactions. User Satisfaction is assigned a weight of 20%, as feedback from users provides crucial insights into the overall quality of the user experience. Finally, Relevance, with a weight of 15%, evaluates how well the responses align with the user's intent and context, ensuring that the interaction is meaningful and appropriate. These metrics collectively assess the system's effectiveness in providing accurate, efficient, and user-friendly customer service.

4.2 Result

The result is shown in the Table 1. The experiment results reveals several key insights into the performance of the intelligent customer service system. Based on the collected data, **Accuracy** consistently scored high across all users, with an average of 94%, indicating that the system is generally successful in providing correct responses. However, there was some variation in **Response Time**, ranging from 2.0 to 3.2 seconds, which suggests that while the system is relatively quick, there is room for improvement in ensuring consistent speed under varying conditions. **User Satisfaction** also remained positive, with an average rating of 4.4 out of 5, reflecting strong user approval of the system's overall performance. This aligns with the high **Relevance** scores, averaging around 89%, which suggests that the system is successful in understanding and addressing the users' specific needs. Overall, the experiment highlights the system's strengths in delivering accurate and contextually appropriate responses, but it also points to potential areas for further optimization, particularly in response time consistency.

Table 1. The Experiment Result

	User	Accuracy (%)	Response Time (s)	User Satisfaction (1-5)	Relevance(%)
1	**User 1**	95	2.5	4.5	90
2	**User 2**	92	3.0	4.2	88
3	**User 3**	98	2.0	4.8	94
4	**User 4**	90	3.2	4.0	85
5	**User 5**	94	2.8	4.4	89

5. Conclusion

In conclusion, this study demonstrates the effectiveness of a modular architecture for intelligent customer service systems powered by Large Language Models (LLMs). The system's performance across key metrics—accuracy, response time, user satisfaction, and relevance—indicates a robust capability in delivering correct, timely, and contextually appropriate responses. The high accuracy and relevance scores confirm that the system meets the user's expectations in terms of providing accurate and meaningful information. User satisfaction further reinforces the system's effectiveness, reflecting positive feedback and engagement. However, the

analysis of response times highlights areas for potential optimization, especially in achieving more consistent performance across different interactions. Overall, the modular approach proves to be a promising framework for developing intelligent customer service systems, with clear avenues for further refinement and enhancement to achieve even better user experiences and operational efficiency.

References

[1] Gupta A. E-Commerce: Role of E-Commerce in today's business[J]. International Journal of Computing and Corporate Research, 2014, 4(1): 1-8.

[2] Hadi M U, Qureshi R, Shah A, et al. A survey on large language models: Applications, challenges, limitations, and practical usage[J]. Authorea Preprints, 2023, 3.

[3] Zhang C, Jin M, Yu Q, et al. Goal-guided generative prompt injection attack on large language models[J]. arXiv preprint arXiv:2404.07234, 2024.

[4] Vahidnia M H. Empowering geoportals HCI with task-oriented chatbots through NLP and deep transfer learning[J]. Big Earth Data, 2024, 8(4): 608-648.

[5] Zhu X, Xue H, Zhao Z, et al. LLM as GNN: Graph Vocabulary Learning for Text-Attributed Graph Foundation Models[J]. arXiv preprint arXiv:2503.03313, 2025.

[6] Miao J, Thongprayoon C, Suppadungsuk S, et al. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications[J]. Medicina, 2024, 60(3): 445.

[7] Rong J, Vu H Q, Law R, et al. A behavioral analysis of web sharers and browsers in Hong Kong using targeted association rule mining[J]. Tourism Management, 2012, 33(4): 731-740.

[8] Ahn J, Verma R, Lou R, et al. Large language models for mathematical reasoning: Progresses and challenges[J]. arXiv preprint arXiv:2402.00157, 2024.

[9] Inavolu S M. Exploring AI-driven customer service: Evolution, architectures, opportunities, challenges and future directions[J]. International Journal of Engineering and Advanced Technology, 2024, 13(3): 156-163.